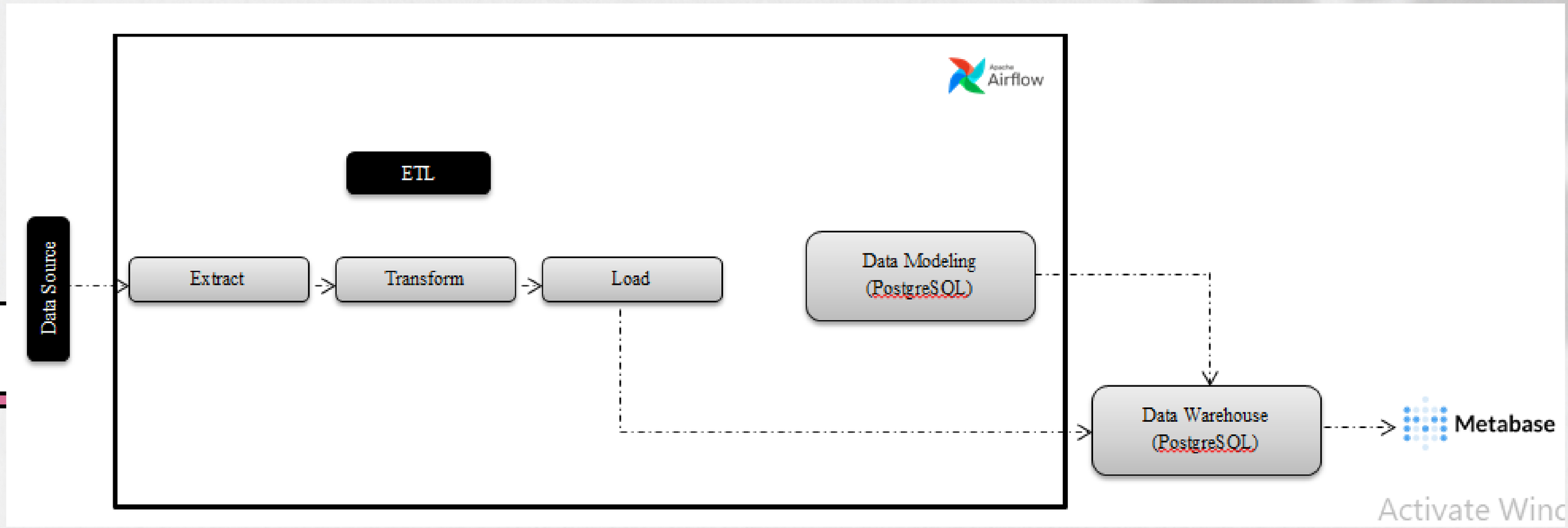
FINAL PROJECT

Presented by:
Kelompok 12
Andi Dhiya Rachmy Ariffia
Haidar Fadhila Fiqa
Meiyra Istiqomah
Dita Artiningtyas

Mentor Pendamping: Ardhi Wahyudi

ARSITEKTUR



ETL

ETL stands for Extract, Transform, Load, and it refers to a process in data management and data warehousing.

Extract

Retrieves and verifies data from various sources

Transform

Processes and organizes extracted data so it is usable

Load

Moves transformed data to a data repository

PROSES ETL

Extract Function dan Transform

```python
# Step 2: Extract Function
def extract ():
  path ="/opt/airflow/data"
  onlyfiles = [f for f in listdir(path) if isfile(join(path, f))]

  # now do the for loop
  for  i in onlyfiles:
    #customer (csv)
    if ".csv" in i and "customer" in i:
      df_customer = pd.DataFrame()
      for j in range(10):
        df = pd.read_csv(path+"/"+i)
        df_customer = pd.concat([df_customer,df])

    #product dan product category (xls)
    if ".xls" in i and "product" in i:
      if "product_category" in i:
        df_product_category = pd.read_excel(path+"/"+i)
      else:
        df_product = pd.read_excel(path+"/"+i)

```

The extract function reads data from different file formats (CSV, XLS, JSON, Parquet, Avro) located in the specified directory (/opt/airflow/data). The extracted data is then returned as a list of Pandas DataFrames.

PROSES ETL

Import Modul

```python
# Step 1: Importing Modules
from airflow import DAG
from datetime import timedelta, datetime
from airflow.operators.python_operator import PythonOperator
import pandas as pd
import fastavro
import psycopg2
from os import listdir
from os.path import isfile, join
from sqlalchemy import create_engine
```

PROSES ETL

LOAD FUNCTION

```python
# Step 3: Load Function
def load(*args, **kwargs):
    connection = psycopg2.connect(
        host="dataeng-warehouse-postgres",
        port=5432,
        dbname="data_warehouse",
        user="user",
        password="password"
    )

    cursor = connection.cursor()
    list_tabel = extract()
    tabel = {}
    tabel['customers'] = list_tabel[0]
    tabel['products'] = list_tabel[1]
    tabel['product_categories'] = list_tabel[2]
    tabel['suppliers'] = list_tabel[3]
    tabel['login_attempt_history'] = list_tabel[4]
    tabel['coupons'] = list_tabel[5]
    tabel['orders'] = list_tabel[6]
    tabel['order_items'] = list_tabel[7]

    engine = create_engine('postgresql://user:password@dataeng-warehouse-postgres:5432/data_warehouse')
```

The load function establishes a connection to a PostgreSQL database (data_warehouse) and loads the extracted data into corresponding tables.

PROSES ETL

CREATING TASKS

```python
# Step 5: Creating Tasks
task_1 = PythonOperator(
    task_id='task_1',
    python_callable=extract,
    provide_context=True,
    dag=dag
)

task_2 = PythonOperator(
    task_id='task_2',
    python_callable=load,
    provide_context=True,
    dag=dag
)
```

Two PythonOperator tasks (task_1 and task_2) are created. task_1 calls the extract function, and task_2 calls the load function.

SETTING UP DEPENDENCIES

```python
# Step 6: Setting up Dependencies
task_1 >> task_2
```

This step defines the dependency between task_1 and task_2, indicating that task_2 should only run after task_1 has completed successfully.

DATA MODELLING

Data modelling digunakan untuk menyusun dan menyimpan data dalam bentuk yang memfasilitasi analisis bisnis yang cepat dan efisien

dim_product_supplier

Terdiri dari table :

1. Product
2. Product Category
3. Supplier

dim_order_coupon

Terdiri dari table :

1. Order
2. OrderItem
3. Coupons

dim_orders

Terdiri dari table Orders

dim_customers_login

Terdiri dari table :

1. Customers
2. Login attempt history

DATA MODELLING

Fact Table Login History

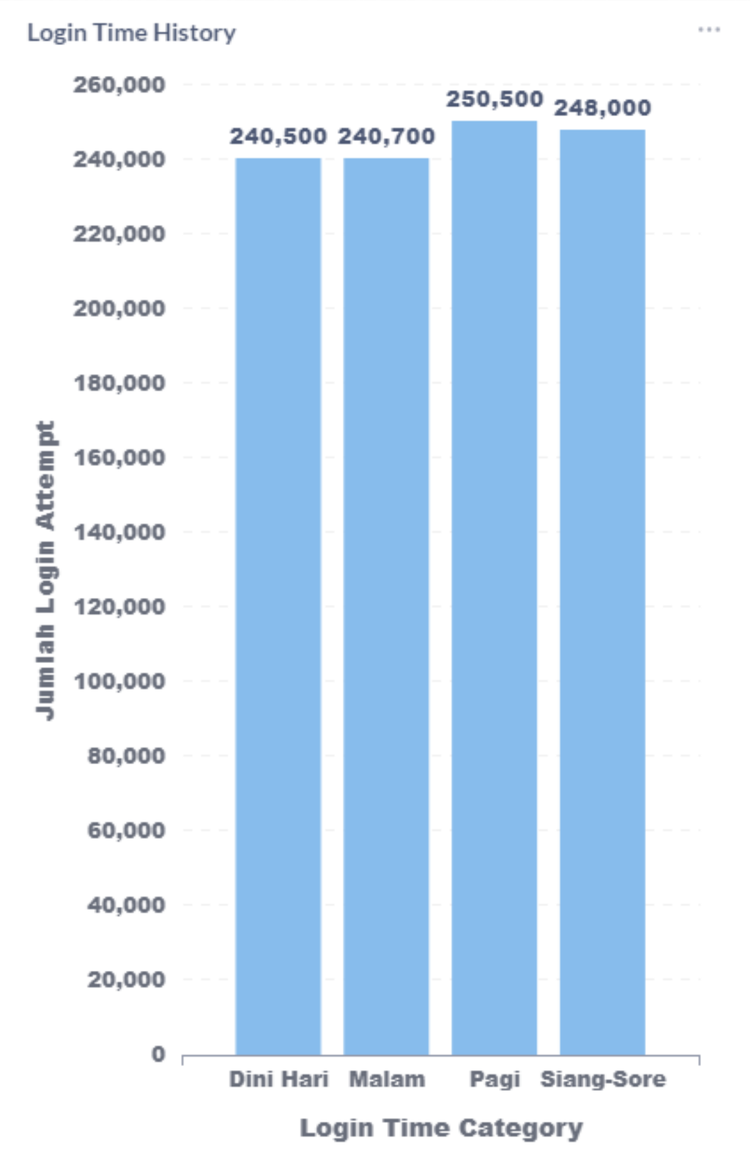
Fact table tersebut berisi informasi kapan customer melakukan log in, dikelompokkan sebagai berikut :
a. Dini Hari
b. Pagi
c. Siang - Sore
d. Malam

Fact Subtotal Diskon

Fact table tersebut berisi informasi total biaya yang dikeluarkan oleh customer setelah memperhitungkan jumlah diskon yang didapatkan

Fact Table Supplier based on Product Category

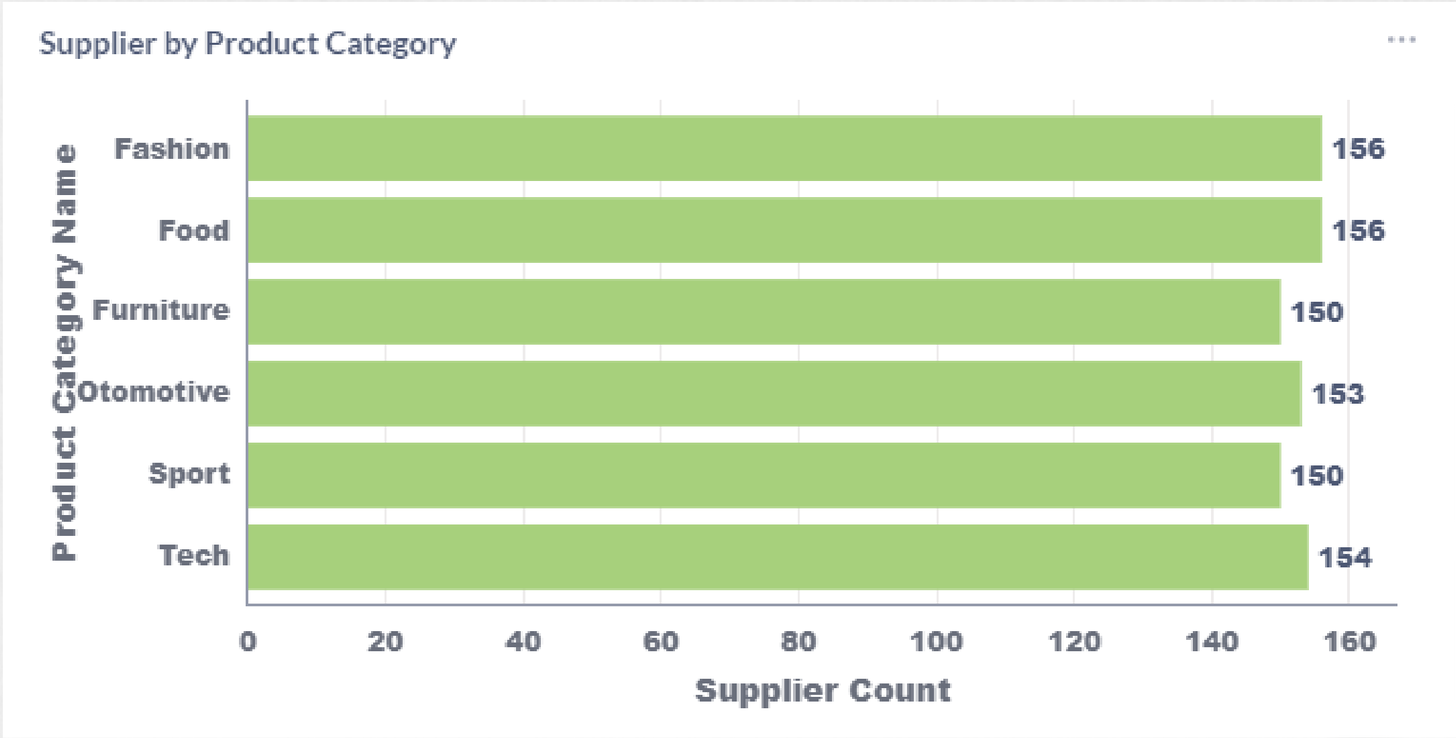
Fact table tersebut berisi informasi terkait jumlah supplier untuk setiap product category



VISUALISASI

Berdasarkan hasil visualisasi, dapat diketahui bahwa customer banyak melakukan login pada pagi hari, yaitu sebanyak 250.500 customer.

VISUALISASI



Berdasarkan hasil visualisasi, dapat diketahui bahwa kategori “Fashion” dan “Food” merupakan product category yang memiliki supplier paling banyak, yaitu sebanyak 156 suppliers.

VISUALISASI

Regional Distribution of Suppliers

Country	Count
Albania	1
Algeria	1
American Samoa	1
Antarctica (the territory South of 60 deg S)	2
Antigua and Barbuda	1
Australia	1
Austria	1
Azerbaijan	1
Bahrain	1

Visualisasi diatas merupakan visualisasi dalam bentuk table yang berisi tentang persebaran wilayah para supplier, yang mana sebagian besar terdiri 1 supplier setiap daerahnya. Namun, ada beberapa daerah yang juga memiliki banyak supplier, misalnya seperti Antarctica yang memiliki 2 supplier dan Timor Leste memiliki 3 supplier.

VISUALISASI

Berdasarkan hasil visualisasi, dapat diketahui rata rata diskon setiap harinya. Terlihat bahwa pada 11 Februari 2023 memiliki rata-rata diskon yang paling besar, yaitu sebesar 54%

SARAN BISNIS

1. Optimalkan pelayanan pada pagi hari. Manfaatkan peluang dengan menyediakan pelayanan atau penawaran spesial pada pagi hari untuk meningkatkan keterlibatan pelanggan.
2. Pertimbangkan strategi pemasaran untuk kategori produk unggulan pada kategori "Fashion" dan "Food" yang memiliki jumlah supplier paling banyak. Evaluasi dan tingkatkan kerjasama dengan supplier di kategori ini.
3. Analisis lebih lanjut tentang wilayah dengan banyak supplier. Fokuskan analisis lebih lanjut pada wilayah-wilayah yang memiliki lebih dari satu supplier untuk mengidentifikasi peluang bisnis atau potensi peningkatan efisiensi operasional

KESIMPULAN

Project yang dilakukan telah berhasil menggabungkan proses ETL, pemodelan data dan analisis visualisasi dengan efektif. Pada pemodelan data memberikan dasar yang kuat untuk analisis, sementara visualisasi dengan metabase bisa memberikan penjelasan yang jelas


THANK
YOU